IDENTIFYING NEW DRUG TARGETS TO COMBAT PATHOGENIC INFECTIONS: AN INTERDISCIPLINARY APPROACH

Ilse Smets [*] **Astrid Cappuyns** [*] **Kristel Bernaerts** [*]
Nadja Van Boxel [**] **Kathleen Sonck** [**]
Sigrid De Keersmaecker [**] **Pieter Monsieurs** [***]
Tim Van den Bulcke [***] **Kathleen Marchal** [***]
Janick Mathys [***] **Bart De Moor** [***]
Jos Vanderleyden [**] **and Jan Van Impe** [*]

[*] *BioTeC–Katholieke Universiteit Leuven,*
W. de Croylaan 46, B-3001 Leuven (Belgium)
Tel: +32-16-32.14.66 Fax: +32-16-32.19.60
e-mail: jan.vanimpe@cit.kuleuven.ac.be
[**] *CMPG–Katholieke Universiteit Leuven,*
Kasteelpark Arenberg 20, B-3001 Leuven (Belgium)
[***] *ESAT SISTA/COSIC – Katholieke Universiteit Leuven,*
Kasteelpark Arenberg 10, B-3001 Leuven (Belgium)

Abstract: Due to the abundant and often inappropriate use of antibiotics, todays' medical treatments are faced with alarming resistance development of pathogenic bacteria. The development of a novel class of antibiotics has therefore become a major research theme. This paper presents a conceptual overview of how this quest is tackled in a multidisciplinary fashion when the focus lies on detecting and understanding regulatory pathways that lead to virulence. The importance of well designed and controlled bioreactor experiments as well as the integration (into mathematical models) of data, collected at different levels and from different sources, will be stressed.

Keywords: biomedical systems, biocontrol, mathematical models, optimal experiment design

1. INTRODUCTION

The acquirement of resistance of pathogenic bacteria to common antibiotics and the development of multidrug resistant strains is raising alarms in health care and fueling demand for new antibiotics. Despite the apparent need for new and effective antibiotics, few novel drug targets have been identified and a very limited amount of new (classes of) antibiotics has been introduced in the last 20 years. Moreover, antibiotics have a broad range effect, killing also the beneficial intestinal microflora. Hence, more sustainable approaches to cope with these infectious bacteria are needed. Efforts are already made by several research groups to come up with alternative ways to combat bacterial infections. The prophylactic and therapeutic use of probiotics can be situated in this context with a boom in functional food R&D activity as a result.

Within this context, the primary goal of this research is to gain insight into the regulatory networks of gene expression in *Salmonella typhimurium*.

The multidisciplinarity of the here presented research lies in the combination of (*i*) well designed and highly controlled bioreactor experiments and (*ii*) integration of data collected at different levels and from different sources into mathematical models. The structure of the paper is, therefore, as follows. First the general aim and main strategy to reach that aim are sketched in Section 2. The subsequent sections elaborate on the different aspects of the strategy.

First, the different types of data and the information that can be inferred from them are introduced in Section 3. Afterwards, the growth of the pathogen and its presumed pathogenesis triggering metabolite production are modelled in Section 4. Hereto, techniques of optimal experiment design will be employed and controlled bioreactor experiments are performed. Finally, genetic network inference is briefly explained as future task in Section 5.

2. GENERAL AIM AND STRATEGY

One of the key factors to explain why the current era is dominated by all sorts of *omics* is the tremendous advancement in measurement techniques of *products* at the intracellular level. As for *genomics*, i.e., the branch of genetics that studies organisms in terms of their genomes (their full DNA sequences), the development of *microarrays* has been a significant milestone.

At the next level (seen from the DNA-(m)RNA-protein perspective), *proteomics*, i.e., the analysis of biological processes by the systematic analysis of a large number of expressed proteins, is becoming an increasingly important research domain. Since the regulatory activity of lots of proteins is enabled or disabled by phosphorylation, the detection of (the evolution of) this phosphorylation is of prime importance.

This type of measurements is however not cheap. Taking samples at the *appropriate* moment, i.e., when something is about to happen or has just happened, saves a lot on the research budget.

Therefore, not only the experiments but also the sampling instances have to be carefully designed. Since the complexity and possible interactions of the underlying biochemical pathways preclude the inference of the cellular behavior merely based on experimental observations, mathematical models are needed. If growth of the microorganism and production of the triggering metabolite can be captured by some mathematical relationships, e.g., macroscopic balance type models, these models can serve as a basis for optimal experiment design studies to ensure experimental data sets with a rich information content.

Once *informative* microarray and (phospho)proteomics data are gathered (i.e., before and after a certain event in order to distinguish between genes that are switched on or off and proteins that become phosphorylated or not), the regulatory genetic network has to be inferred. Hereto, recently developed bioinformatics tools are employed.

The above research aspects will be discussed more extensively in the following sections.

3. INFORMATION AT DIFFERENT LEVELS

3.1 Microarray data

Microarray experiments measure the expression level of many genes simultaneously and can therefore be considered as upscaled Northern hybridizations. Each spot on the array represents a distinct coding sequence of the genome of interest. The spots (probes) typically consist of PCR-amplified cDNAs of approximately 300 bp. During a microarray experiment, mRNA of a reference and induced sample are isolated, reverse transcribed into cDNA, and labeled with distinct fluorochromes. Subsequently, both cDNA samples are hybridized simultaneously to the array. Fluorescent signals of both channels are measured and used for further analysis (for more extensive reviews on microarrays reference is made to (Brown and Botstein, 1999; Blohm and Guiseppi-Elie, 2001; Southern, 2001)).

3.2 (Phospho)proteomics data

Microarrays are useful to detect the changes of up and down regulated genes, but disregard alterations at protein level. In some cases, the correlation between mRNA and protein level (activity) can expected to be small due to the presence of phenomena not visible at mRNA level. The nature of these phenomena can be elucidated using a proteomics approach.

Proteomics can be defined as the identification, characterization and relative quantification of all proteins involved in a particular pathway, organelle, cell, tissue, organ or organism that can be studied in concert to provide accurate and comprehensive data about that system. Proteomics originates from high-resolution two-dimensional gel electrophoresis (2DE) for protein separation and quantification. Today, mass spectrometry (MALDI-TOF-MS) is by far the most common used method for protein identification from 2D gels. By peptide-mass fingerprinting (PMF) an experimental profile of peptide masses (i.e., a protein separated by 2D, and digested with a protease) can be compared to a profile theoretically

calculated from the known sequences in a non-redundant protein database (Blackstock, 2000).

Posttranslational modification of proteins is a key regulatory event in many cellular processes including recognition, signaling, targeting and metabolism. In general, posttranslational modifications serve as on-off switches or modulators of protein activity and targeting and also regulate the assembly and disassembly of macromolecular complexes including protein-ligand, protein-protein and protein-nucleic acid interactions. Reversible posttranslational modification of proteins includes the covalent attachment or removal of a functional group. Many key regulatory proteins in the cell are always present and they are not up or down regulated by gene-expression control. Their activity often depends on posttranslational modification, and therefore their activity is not truly reflected by protein or RNA-expression analysis (Jensen, 2000). Phosphoproteomics is an obvious choice for detecting reversible protein phosphorylation events in the function of time, as protein phosphorylation is the major regulator of important cell-signaling processes. There are several methods to investigate quantitative changes in protein phosphorylation in complex protein mixtures. A remarkable breakthrough was proposed by (Zhou *et al.*, 2001). The approach consists of three steps: (*i*) selective phosphopeptide isolation from a peptide mixture via a cascade of chemical reactions, (*ii*) phosphopeptide analysis by a combination of automated liquid chromatography and mass spectrometry (LC-MS-MS), and (*iii*) identification of the phosphoprotein and the phosphorylated residue(s) by correlation of tandem mass spectrometric data with sequence databases. Another method uses 2DE separation of protein samples, followed by Western blotting with antibodies against phosphorylated amino acids (antiphosphotyrosine, antiphosphoserine and antiphosphothreonine). Phosphorylated proteins are subsequently identified by mass spectrometry.

4. MACROSCOPIC MODELLING AND OPTIMAL EXPERIMENT DESIGN

If the bacterial pathogenic response is triggered by a certain metabolite, then it is evident that (*i*) the reaction network or mechanism that produces the metabolite as well as (*ii*) the *downstream* reactions that this metabolite initiates are both possible drug targets, once clearly understood. While most of the reported studies in this context rely on (batch-wise) test tube or erlenmeyer experiments, a controlled environment and possibly fed-batch or continuous type experiments are a prerequisite to clearly distinguish the phenomena that potentially influence the studied process. If for example the influence of a certain carbon source is to be tested, then the pH has to be controlled since the catabolic reactions following the consumption of the carbon source could influence the pH, hence, hampering the distinction between both phenomena.

To enhance this understanding, first of all, experiments have to be designed from which the production mechanism of the metabolite can be inferred. In a second step this production must be stimulated such that information rich data can be collected to unravel the pathways that are triggered by the (abundant) presence of the metabolite.

To get acquainted with the microbial growth and production process, some preliminary batch experiments have been performed.

Experimental conditions. The studied bacterial species is the pathogen *Salmonella typhimurium*. Batch cultures were conducted in a computer controlled BioFlo 3000 benchtop fermentor (New Brunswick Scientific, USA) with an autoclavable vessel of 5 L working volume. An overnight preculture was transferred to the fermentor vessel containing 4.0 L Luria-Bertani medium. PID cascade controllers ensured that the fermentation temperature as well as the pH and the dissolved oxygen (DO) were kept constant as to mimic the human intestinal environment. Glucose is provided as the sole external carbon source.

Measurements. Culture media samples are removed at regular intervals. CFU [1] /mL values are obtained by plate counting. Glucose concentrations are determined using an enzymatic test kit while the metabolite concentration is established by a specific bioassay.

Mathematical tools. The implemented identification routine for model parameter estimation is the `e04UCF` routine from the NAG library (Numerical Algorithms Group) in Fortran. Apart from Fortran, Matlab 6.1 (The Mathworks Inc., Natick) is used as simulation software.

Optimal experiment design. When modelling growth or production kinetics, the first issue is the selection of an appropriate model structure. Once the structure has been determined, a *unique* solution for the set of corresponding model parameters (which have to be estimated from experimental data) has to be found. A unique identification of the parameter set is only possible if the available data are sufficiently rich. In system identification theory this is known as *persistent excitation* of the system. Hence, it is clear that an efficient

[1] colony forming units

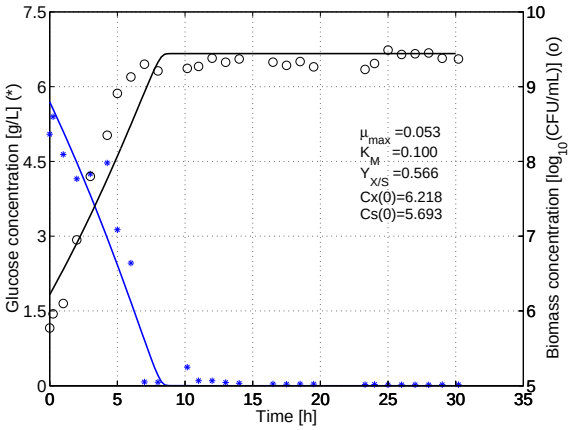

Fig. 1. Evolution of the glucose (*) and biomass (o) concentration in function of time.

experimental planning plays a crucial role in the practical identifiability of the kinetic parameters.

Figure 1 depicts the evolution of the glucose substrate (stars) and biomass concentration (in CFU/mL, bullets) in function of time during a first preliminary batch experiment.

When focusing on the growth phase, a simple Monod model (Equation 2) seems appropriate. The link between the specific growth rate and the substrate consumption is the so-called *linear law* in which, for the time being, the maintenance term is neglected. All substrate consumed is therefore assumed to be built in as new biomass with a certain efficiency or yield factor $Y_{X/S}$ [10^3g/CFU]. The evolution in time of the substrate concentration C_S [g/L] (i.e., glucose) and the biomass concentration C_X [CFU/mL] is then described by following system of mass balance equations:

$$\begin{aligned} \frac{dC_S}{dt} &= -\frac{\mu}{Y_{X/S}} \cdot C_X \\ \frac{dC_X}{dt} &= \mu \cdot C_X \end{aligned} \tag{1}$$

in which

$$\mu = \mu_{max} \frac{C_S}{C_S + K_M}. \tag{2}$$

In this specific growth rate expression, μ_{max} [1/h] is the maximum specific growth rate and K_M [g/L] the half saturation constant.

However, correct identification of the parameters is not a trivial task since (*i*) the experimental data points are scarce and (*ii*) batch experiments are known as not the most optimal setup for estimation of both Monod constants at once (Holmberg and Ranta, 1982). It has been proved that the extension of the batch experiment by a feeding phase with time-varying feed rate leads to a higher accuracy of the parameter estimates. In this context, the following conjecture was formulated by (Van Impe and Bastin, 1995):

> A feed rate strategy which is optimal in the sense of process performance is an excellent starting point with respect to estimation of those parameters with large influence upon process performance.

With biomass growth optimization in mind, optimal limiting substrate feed rate profiles are often of the bang-singular-bang type (Van Impe and Bastin, 1995) with a first maximum feeding or batch phase, followed by a singular phase (during which the substrate concentration is kept constant) and ending with a batch phase until all available substrate is consumed. Therefore, such a profile is proposed as starting point for unique parameter estimation by means of optimal experiment design techniques (see also, e.g., (Versyck *et al.*, 1997)).

Parameter estimation can be formulated as minimization of the following *identification functional* $\mathcal{J}$ by optimal choice of the parameter vector $\mathbf{p}$:

$$\mathcal{J} \triangleq \int_0^{t_f} (\mathbf{y}(\mathbf{p}) - \mathbf{y}_\mathrm{m})^T \mathbf{Q} (\mathbf{y}(\mathbf{p}) - \mathbf{y}_\mathrm{m})\, dt \tag{3}$$

in which $\mathbf{y}_\mathrm{m}$ is the vector of measured outputs, $\mathbf{y}(\mathbf{p})$ is the vector of model predictions by using the parameter vector $\mathbf{p}$, and $\mathbf{Q}$ is a user-supplied square weighting matrix. To analyze and quantify the information content of the state trajectories obtained in a certain experiment, the *Fisher information matrix* can be called upon:

$$\mathcal{F} \triangleq \int_0^{t_f} \left(\frac{\partial \mathbf{y}}{\partial \mathbf{p}}\right)^T \mathbf{Q} \left(\frac{\partial \mathbf{y}}{\partial \mathbf{p}}\right) dt \tag{4}$$

$\mathbf{Q}$ is normally selected as the inverse of the measurement error covariance matrix. This choice of the weighting matrix $\mathbf{Q}$ implies that the more a measurement is corrupted by noise, the less it will count in the information criterion. Depending on the requirements imposed by the application, a specific scalar function of this Fisher information matrix is used as the performance index for optimal experiment design to enhance the parameter identifiability. In this study, the following so-called *modified E-criterion* is adopted:

$$\Lambda(\mathcal{F}) = \frac{\lambda_{max}(\mathcal{F})}{\lambda_{min}(\mathcal{F})} \tag{5}$$

which represents the ratio of the largest to the smallest eigenvalue of $\mathcal{F}$. To enhance parameter identifiability this condition number should approximate one as to induce circular lines of constant functional values and a conelike functional shape of $\mathcal{J}$.

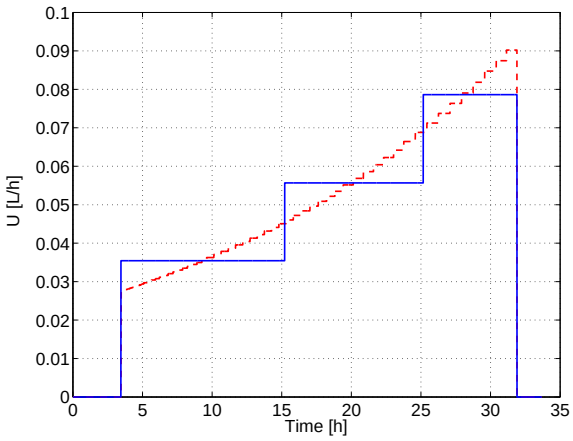

Fig. 2. Optimal and suboptimal feeding rate profile for growth parameter identification.

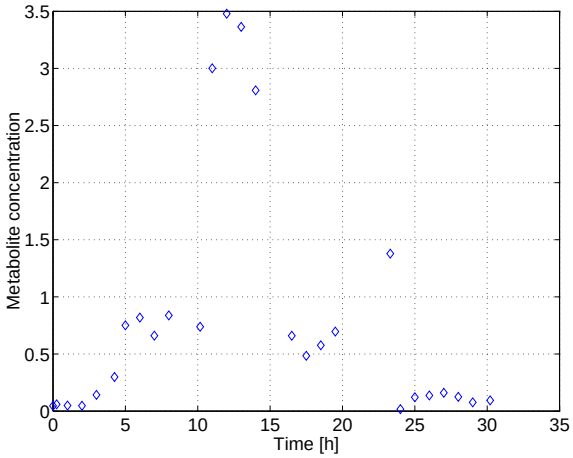

Fig. 3. Evolution of the triggering metabolite concentration in function of time.

Following the above procedure, the optimal feed rate profile U [L/h] for glucose is as depicted in Figure 2 (dotted line): a first batch phase followed by a singular feeding phase ending in a short batch phase. The condition number Λ is 36.027 and the optimal singular substrate concentration C_S^* equals 0.361 g/L. A lower condition number can be obtained when the substrate concentration during the singular phase is allowed to be lower, which is however hard to realize in practice. To further simplify the practical aspects of the proposed experiments, one could try to approximate the singular *time-varying* feeding phase by stepwise increasing the feed rate during that period (while the total added volume during that phase remains the same). Figure 2 (solid line) illustrates the obtained suboptimal profile for which the condition number Λ is now 46.995.

Production modelling. The evolution of the metabolite production can be seen in Figure 3.

Unfortunately, it is impossible to derive the metabolite production mechanism from this profile Future experiments (and/or other measurement protocols for the metabolite) will have to elucidate the metabolite production mechanism.

Once a simple, though reliable macroscopic balance type mode for growth and metabolite production has been derived, this model will be exploited in optimal experiment design studies to establish the most informative experiments with respect to specific gene up- or down regulation.

5. REGULATORY NETWORK INFERENCE

With the information obtained from the previous step, genetic networks can be inferred. The most promising implementations of network inference are based on Bayesian networks. Bayesian networks are probabilistic models consisting of a graphical structure and conditional probabilities. A Bayesian network allows both a compact representation of the joint probability distribution over a large number of variables, and an efficient way of using this representation for statistical inference. It consists of a directed acyclic graph that models the interdependencies between the variables, and a conditional probability distribution for each node with incoming edges. Given a graph, it is possible to learn the probability distributions from the available data and Bayesian priors. One then searches in the space of candidate graphs for a graph that models the dependencies in the data best. In the context of genetic network inference, the nodes in the network represent the expression levels of the genes (variables) and the edges correspond to the interactions. Bayesian (belief network) are an almost natural choice to model regulatory pathways. Since biological networks are structured hierarchically, connections between genes are sparse. In a Bayesian network such sparse connections can easily be represented by the conditional independencies. Since Bayesian networks are probabilistic in nature they can capture the stochasticity (either biological or experimental) of the system. Moreover, Bayesian networks can cope with the presence of unobserved values (hidden variables, e.g., not measured protein-protein interactions). The graphical representation reflects the real biological structure and this structure can be inferred independently from the parameter estimation (maximum a posteriori, Monte-Carlo sampling). Most important is probably the natural way by which prior information can be introduced into the model.

This prior knowledge is gathered via the construction of a knowledge base system or ontology. An ontology is a specification of a *conceptualization*, designed to describe and store domain knowledge, external to the system (Gruber, 1993).

A conceptualization is an abstract, simplified view of the objects or concepts, that are assumed to

exist in a given research domain and the relations that hold among them (Karp *et al.*, 2000). This set of concepts and their relations is reflected in the representational vocabulary that a knowledge base uses to represent knowledge. An ontology in the molecular-biology field will typically contain concepts such as *gene*, *gene name*, *protein*, *molecular function*, *biological process* etc. Relations between these concepts can be defined. For instance, the concept *protein participates in* (relation) a certain *biological process* (related concept).

6. CONCLUSIONS

Within the context of the development of a new class of antibiotics, this paper introduces a multidisciplinary approach to unravel the regulatory networks of gene expression in *Salmonella typhimurium*. The key factors in this search are, according to the authors, well designed and controlled bioreactor experiments as well as the integration (into mathematical models) of data, collected at different levels and from different sources. Although of preliminary nature, the results obtained thus far are promising.

ACKNOWLEDGMENTS

Ilse Smets and Kristel Bernaerts are postdoctoral fellows with the Fund for Scientific Research Flanders (FWO). Work supported in part by (*i*) the IWT-GBOU-20160 project, (*ii*) the OT/99/24 and OT/03/30 projects of the Research Council of the Katholieke Universiteit Leuven and (*iii*) the Belgian Program on Interuniversity Poles of Attraction, initiated by the Belgian State, Prime Minister's Office for Science, Technology and Culture. The scientific responsibility is assumed by its authors.

REFERENCES

Blackstock, W. (2000). Trends in automation and mass spectometry for proteomics. In: *Proteomics: A Trends Guide* (W. Blackstock and M. Mann, Eds.). pp. 12–17. Elsevier Science. London.

Blohm, D.H. and A. Guiseppi-Elie (2001). New developments in microarray technology. *Current Opinion in Biotechnology* **12**, 41–47.

Brown, P.O. and D. Botstein (1999). Exploring the new world of the genome with DNA microarrays. *Nature Genetics* **21**, 33–37.

Gruber, T.R. (1993). A translation approach to portable ontologies. *Knowledge Acquisition* **5**(2), 199–220.

Holmberg, A. and R. Ranta (1982). Procedures for parameter and state estimation of microbial growth process models. *Automatica* **18**(2), 281–193.

Jensen, O.E. (2000). Modification-specific proteomics: systematic strategies for analyzing post-translationally modified proteins. *Proteomics: A Trends Guide* **July**, 36–42.

Karp, P.D., M. Riley, M. Saier, I.T. Paulsen, S.M. Paley and A. Pellegrini-Toole (2000). EcoCyc and MetaCyc databases. *Nucleic Acids Research* **28**, 56–59.

Southern, E.M. (2001). DNA microarrays. history and overview. *History and overview Methods Molecular Biology* **170**, 1–15.

Van Impe, J.F. and G. Bastin (1995). Optimal adaptive control of fed-batch fermentation processes. *Control Engineering Practice* **3**(7), 939–954.

Versyck, K.J., J.F. Claes and J.F. Van Impe (1997). Practical identification of unstructured growth kinetics by application of optimal experimental design. *Biotechnology Progress* **13**(5), 524–531.

Zhou, H.L., J.D. Watts and R. Aebersold (2001). A systematic approach to the analysis of protein phosphorylation. *Nature Biotechnology* **19**, 375–378.